Machine Learning Methods Applied to Building Energy Production and Consumption Prediction

Paulo Lissa[1*], Dayanne Peretti[1], Michael Schukat[1], Enda Barrett[1], Federico Seri[1] and Marcus Keane[1]

[1]College of Science and Engineering, National University of Ireland, Galway, Ireland
*paulo.lissa@nuigalway.ie

Abstract. The utilization of renewable sources of energy is growing all over the world due to pressure for sustainable solutions. It brings benefits to the environment, but also adds complexity to the electricity grid, which faces energy balancing challenges caused by an intermittent production from this kind of generation. Having a good energy prediction is essential to avoid losses and improve the quality and efficiency of the energy systems. There are many machine learning (ML) methods that can be used in these predictions; however, every consumer is different and will behave in a distinct way. Therefore, the objective of this article is to compare the application of different ML methods, aiming to predict PV energy production and energy consumption for residential users. Four different ML methods were applied in a real dataset from the RESPOND project: Linear Regression, Decision Forest regression, Boosted Decision Tree Regression and Neural Network. After the simulation, the predicted values were compared against the real data, considering 150 days of measurement from two Irish houses. Overall, all the algorithms applied achieved mean errors below 14%, but the Boosted Decision Tree overperformed, with mean errors of 2.68% and 10% for energy consumption and energy production prediction, respectively.

Keywords: Machine learning, Energy production, Consumption Prediction.

1 Introduction

Energy generation through renewable sources is becoming popular in recent years. The European Union (EU) targets to achieve about 20% of renewable energy production in 2020 [1] and at least 27% in 2030 [2]. Along with this expected growth, other challenges start to arise, mainly because energy produced from wind or solar sources depends on weather conditions and presents an intermittent capacity. This will tend to increase the variability of overall electricity supply, thus making its integration to the grid a complex process [3].

Understanding and predicting how electricity network works, including distributed generation from renewables, is essential in this new framework, as it can bring benefits to the utilities. Contemporary solutions for energy balance, such as backup fossil

powerplants [3] and storage [9], are costly and sometimes not efficient. With a proper energy management system (EMS) utilities can provide new services. These include Demand Response (DR) solutions, where utilities can give benefits to users that change their consumption behavior according to the network load, hence reducing total energy demand during peak times [19]. Moreover, having information about the grid status can help utilities to plan their own energy production, thus avoiding unnecessary costs with new assets otherwise required to match peak demand over small periods.

To support this new trend, some data-driven methods for energy production and consumption prediction have been arising, ranging from statistical models to complex Machine Learning (ML) algorithms. These aim to find correlations and meaning among variables in large datasets. Although more than 80% of the previous studies about energy consumption prediction have been carried on non-residential customers [4], research from [25] shows that in the EU residential applications correspond to 42% of the total energy flexibility potential, whereas 31% comes from industry and 27% in the tertiary sector.

In summary, the main contribution of this work is to assess Machine Learning techniques for energy prediction and to deploy a simulation environment, aiming to provide the following predictions for a hypothetical EMS:

1. Photovoltaic (PV) energy generation.
2. Residential energy consumption for a small group of houses.

The rest of this paper is organized as follows: *Related work*, which shows the related work regarding energy production and consumption prediction. *Machine Learning* section provides information about the principles and techniques applied in this research. *Environmental Setup* describes how the environment monitoring has been structured and explains the dataset preparation stage. *Results* section presents all the relevant outputs of our experiments, comparing the prediction methods and real data. Finally, *Conclusions and Future works* recaps the main points of the paper, introducing ideas for future work.

2 Background Research

Electricity is a development indicator, it boosts country's economy and brings comfort to our homes, improving quality of life in most of daily tasks. However, it is also strongly associated with CO_2 emissions, where buildings represent 36% of the total produced gas in the EU [5]. As energy production using fossil power plants is one of the CO_2 emissions reasons, the use of renewable sources is increasing, therefore affecting directly the energy matrix. Renewables represented almost two-thirds of new net world electricity capacity extensions in 2016, with almost 165 gigawatts (GW) coming online. Between 2017 and 2022, it is expected that the global renewable electricity capacity is to expand by over 920 GW, an increase of 43% [6].

Amasyali and El-Gohary (2018) carried out an extensive review [4] on data-driven building energy consumption prediction, having categorized more than 60 previous studies across five categories: type of building, temporal granularity, type of energy

consumption, type of data and ML algorithm. As a result, they identified that 19% of models belong to residential buildings and the granularity chosen was mostly hourly (57%) followed by daily (15%). Most of datasets considered only the overall energy consumption (47%) from electricity meter and 67% of the models used real data instead of simulated or public benchmark data. Finally, the most frequent ML algorithms applied were artificial neural networks (ANN) and support-vector machine (SVM), with 47% and 25% respectively.

In a different approach, Naji at al. [7] proposed the application of EML (extreme learning machine) algorithm for estimating energy consumption based on a building envelope's parameters, district heating and cooling loads, achieving an accuracy improvement when comparing the results against genetic programming and artificial neural network. Authors in [8] and [10] utilized a genetic algorithm applied for building performance, the first predicting energy consumption and the second one predicting heating/cooling. Besides ML methods, there is also a physical modelling approach, known as engineering methods or white-box models, but they rely on thermodynamic rules for a detailed energy modelling and analysis [4] and thus are not part of the present work.

Regarding energy production prediction, Das at al. [11] assessed more than 20 recent works about forecasting of PV generation, from physical models to ML, and compared their performance across different factors, such as accuracy, reliability, computational cost and complexity. According to the study, ANN and SVM-based forecasting models performed well under rapid and varying environmental conditions. Voyant at al. [12] presented another list with almost 50 works where ML was applied through different methods, with ANN the most popular followed by SVM, regression trees and others. There is no common agreement with regard to the evaluation criteria, but as a reference the root-mean-square error (RMSE) of some of them ranged from 5% to 24%. [13] compared three different prediction models for a PV plant in south Italy: phenomenological detailed model, Multi-Layer Perceptron (MLP) neural network and a regression approach. The results demonstrated that more accurate predictions can be reached by statistical machine learning approaches.

The authors in [14] applied ML algorithms, such as SVM and Forest regression, in order to predict solar radiation values for seven different places in Spain. Our proposed research is about on PV generation, but there are other important studies that show application of ML for different renewable energy source. For instance, [15] and [16] presented a review of current methods for wind power generation forecasting.

3 Machine Learning

Machine Learning is a subfield of computer science that is classified as an artificial intelligence method. It can be used in several domains and one of the advantages is the capability of solving problems which are impossible to be represented by explicit algorithms. Some of the ML methods are regression based, which can be widely used to create projections about future, with the objective to predict a numeric target.

The best ML model will rely on the equilibrium between predicted error and complexity of the system. Depending on the database particularities, a complex model may result in a greater error than using a simple model, as shown in Fig. 1.

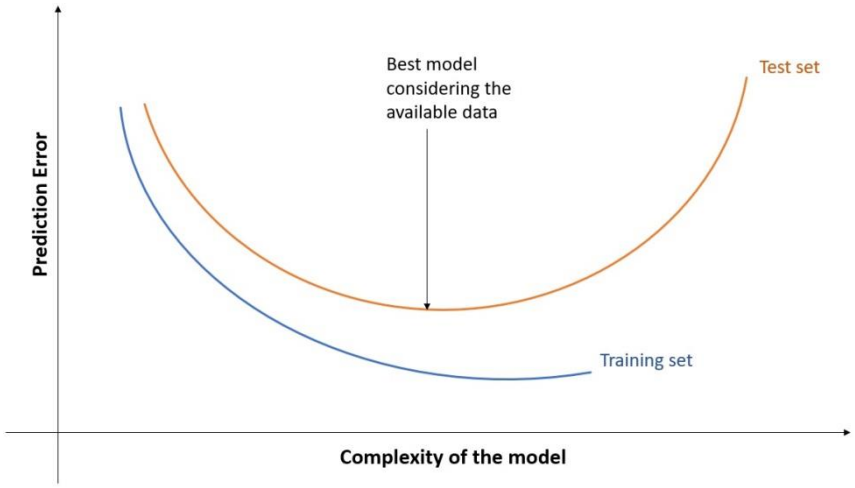

Fig. 1. Model complexity versus prediction error.

All methods to be presented here are regression-based. The following subsections describe the methods applied to our proposed prediction model.

3.1 Linear

Linear regression is a statistical method, which has been adopted for using in ML. Spite of being one of the simplest models for a basic predictive task, this method also tends to work well on high-dimensional sparse datasets [21]. The classic regression problem involves a single independent variable and a dependent variable, this is called simple regression. Multiple linear regression involves two or more independent variables that contribute to a single dependent variable. Problems in which multiple inputs are used to predict a single numeric outcome are also called multivariate linear regression.

3.2 Decision Forest

Decision trees are non-parametric models that perform a sequence of simple tests for each instance, traversing a binary tree data structure until a leaf node (decision) is reached. The advantage of decision trees is that this method is efficient in both computation and memory usage during training and prediction.

Decision Forest model consists of an ensemble of decision trees. Each tree in a regression decision forest outputs a Gaussian distribution as a prediction. An aggregation is performed over the ensemble of trees to find a Gaussian distribution closest to the combined distribution for all trees in the model [22].

3.3 Boosted Decision Tree Regression

Boosting is one of several classic methods for creating ensemble models, along with bagging, random forests, and so forth. In Azure Machine Learning Studio [18], boosted decision trees use an efficient implementation of the MART gradient boosting algorithm, which is a ML technique for regression problems. It builds each regression tree in a stepwise fashion, using a predefined loss function to measure the error in each step and correct for it in the next. Thus, the prediction model is an ensemble of weaker prediction models [20].

3.4 Neural Network Regression

Although neural networks are widely known for applications in deep learning and modeling complex problems, such as image recognition, they are easily adapted to regression problems. Any class of statistical models can be termed a neural network if they use adaptive weights and can approximate non-linear functions of their inputs. Thus, neural network regression is suited to problems where a more traditional regression model cannot fit a solution [23].

The layers of a neural network are made of nodes, the place where computation happens. A node combines input from the data with a set of coefficients, or weights, that either amplify or dampen that input, thereby assigning significance to inputs with regard to the task the algorithm is trying to learn. These input-weight products are summed and then the sum is passed through a node's so-called activation function, to determine whether and to what extent that signal should progress further through the network to affect the ultimate outcome. If the signals pass through, the neuron has been "activated." Fig. 2 shows a diagram of what one node might look like [24].

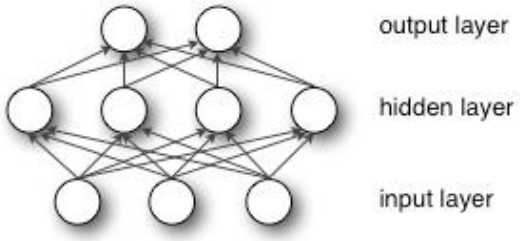

Fig. 2. Neural Network layers.

A node layer is a row of those neuron-like switches that turn on or off as the input is fed through the net. Each layer's output is simultaneously the subsequent layer's input, starting from an initial input layer receiving your data. Pairing the model's adjustable weights with input features is how we assign significance to those features regarding how the neural network classifies and clusters input [24].

4 Environment Setup

The case study chosen is part of the Irish pilot from RESPOND project [17]. It consists of data collected from two houses in the Aran Islands over 150 days, from May to September 2019, both equipped with PV panels. In this experiment, the data was grouped and houses were considered as a cluster because of the goal to analyze energy generation and consumption in the whole grid. The dataset was then uploaded to Microsoft Azure Machine Learning Studio [18], where the information was processed following our proposed architecture showed in Fig. 3.

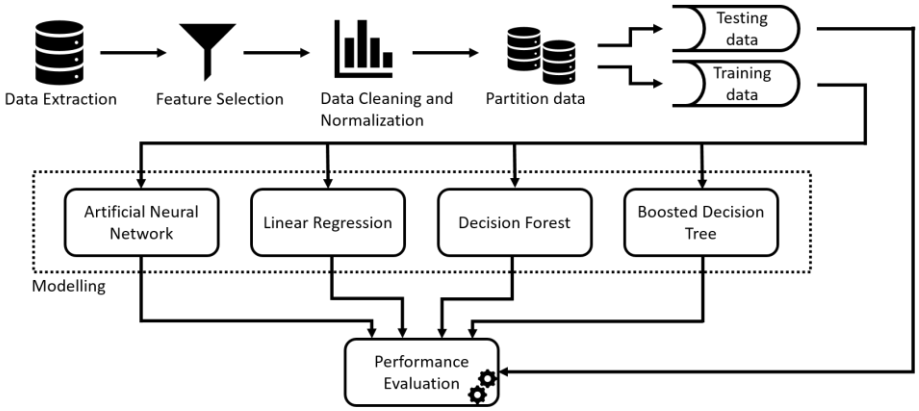

Fig. 3. Process diagram.

4.1 Data Extraction

Data extraction is the first step of the process. It is a collection of all available data from sensors and weather information from external sources. Furthermore, an additional classification features for day classification has been created to improve the algorithm decisions in later stages. The initial dataset has hourly resolution and is composed of the following features:

- Complete timestamp (day, month, year, hour).
- PV generation.
- Energy consumption (electricity meter).
- Weather (temperature, precipitation, humidity, wind speed, solar radiation).
- Day classification (weekend or working days).

The energy consumption data comes from utility's electricity meter, which is the most common source of this kind of measurement. It does not consider social aspects, such as number of family members, because it would be hard to track changes considering large groups, thus making the generalization process complex.

4.2 Feature Selection

This stage is where the selection of features for energy production or consumer consumption prediction model happens. For instance, the production prediction model uses of almost all available variables, only energy consumption is removed from the dataset. On the other hand, consumption prediction does not depend on PV production or some weather features, such as wind speed or solar radiation, so the final dataset is reduced in that case. This practice helps the algorithms to converge faster and allows a better generalization.

4.3 Data Cleaning and Normalization

This process aims to make the dataset as homogeneous as possible. Data cleaning works identifying incomplete, incorrect, inaccurate or irrelevant parts of data and then replaces, modifies or deletes the dirty or coarse data. The normalization process aims to change the values of numeric columns in the dataset to a common scale, without distorting differences among variables values. For example, temperature values range from 0 to 25, while solar radiation can achieve values greater than 500.

The normalization method applied to the proposed experiment is the Z-Score, where the values in the specified columns are transformed using equation 1.

$$z = \frac{x - mean(x)}{stdev(x)} \quad (1)$$

where mean and standard deviation are computed for each column separately.

4.4 Test and Training Dataset

Once data preparation is done and ready for processing, the dataset is randomly divided and follows two different paths:

- 70% of data is sent to training models.
- 30% of data is separated and used for testing purposes, to be compared against trained models later.

4.5 Modelling

The training dataset received from previous stage is trained across four different regression methods: Linear, Decision Forest, Boosted decision tree and Neural Network. Detailed description about each one can be found in the Section 3.

4.6 Performance Evaluation

The predicted results from the different models are compared against the real data. Our experiment considers a performance evaluation of daily and hourly predictions for both, PV generation and energy consumption. Azure ML Studio provides outputs about

overall accuracy, but we have also added a Python script to plot and calculate additional outputs allowing an intuitive visual analysis.

5 Results

Over the performance evaluation stage, the four algorithms have been parameterized in different ways targeting to minimize errors. Azure ML Studio allows us to input a range of parameters to be trained. For instance, instead of using a static value for number of decision trees in the Decision Forest method you can set a range of numbers and the algorithm will try all of them, choosing the best combination of parameters. Immediately below you can find the final parametrization of each method:

- Linear Regression: Method: Ordinary Least Squares. Regularization weight: 0.001.
- Decision Forest Regression: Resampling method: Bagging. Number of trees: 8. Maximum depth of the decision trees: 32. Number of random splits per node: 128.
- Boosted Decision Tree Regression: Maximum leaves per tree: 20. Minimum number of training instances: 10. Learning rate: 0.1. Total number of trees constructed: 100.
- Neural Network regression: Hidden layer specification: Fully connected case. Number of hidden nodes: 100. Learning rate: 0.0001. Number of learning iterations: 100. Initial learning weights diameter: 0.1. Normalizer: Gaussian.

5.1 Daily Predictions

The objective of the model presented in this paper is to provide energy forecast to utilities, in order to help them planning their own energy production in a day-ahead base. The weather inputs for PV production relate to one day before the actual day. The model could also use forecast data from two or three days before, but the accuracy will drop. Users' consumption prediction considers historical trends and also weather forecast.

Table 1. Prediction errors.

	Linear	Decision Forest	B. Decision Tree	Neural Network
PV Production				
MAE	35,24%	27,63%	21,84%	26,33%
Mean Error	14,10%	14,24%	9,99%	11,92%
Energy Consumption				
MAE	16,38%	11,93%	10,31%	12,70%
Mean Error	5,56%	4,31%	2,68%	4,98%

The results from our tests show a mean error below 14% for PV production prediction and below 6% for energy consumption. The mean absolute error (MAE) ranges from 21% to 35% for PV production and from 10% to 16% for energy consumption. The

boosted decision tree presents the best performance across the methods, followed by the neural network. Table 1 presents a compilation of results

In some circumstances, the use of mean error can better represent the reality. For instance, if the utility is analyzing the forecast of a huge number of houses, some of them will present positive errors, predicting more energy than necessary, and others will have the opposite effect with negative errors, so this kind of measurement could result in a better balance than absolute values. Overall, the four methods follow the real data trend, as can be seen in Fig. 4 (PV generation prediction) and Fig. 5 (Energy consumption prediction).

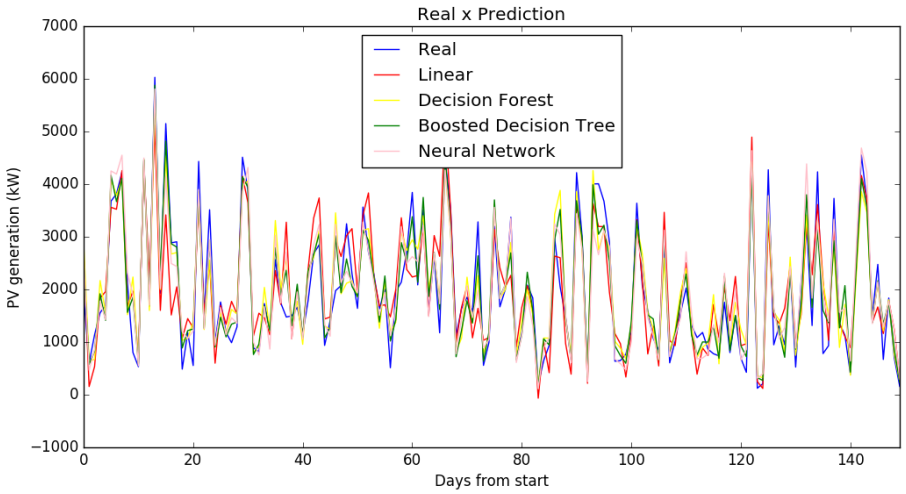

Fig. 4. PV generation prediction model comparison.

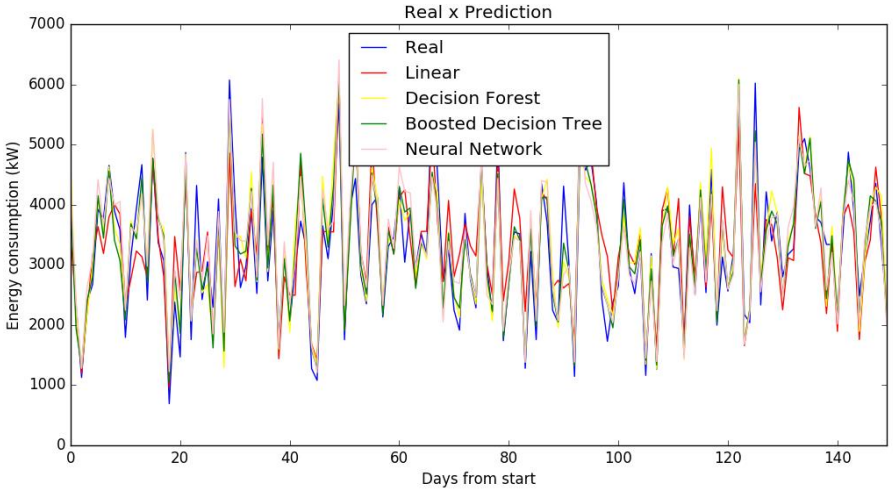

Fig. 5. Energy consumption prediction model comparison.

5.2 Hourly Predictions

Hourly energy consumption prediction can be hard to perform, mostly because consumers' behavior can vary throughout the week with no defined pattern. PV production forecasting can also suffer variation and uncertainty, due to sudden weather changes. In order to show the accuracy across the models, we selected the best performer (boosted decision tree) and the worst (linear regression), both presented in Fig. 6 (consumption prediction) and Fig. 7 (PV generation).

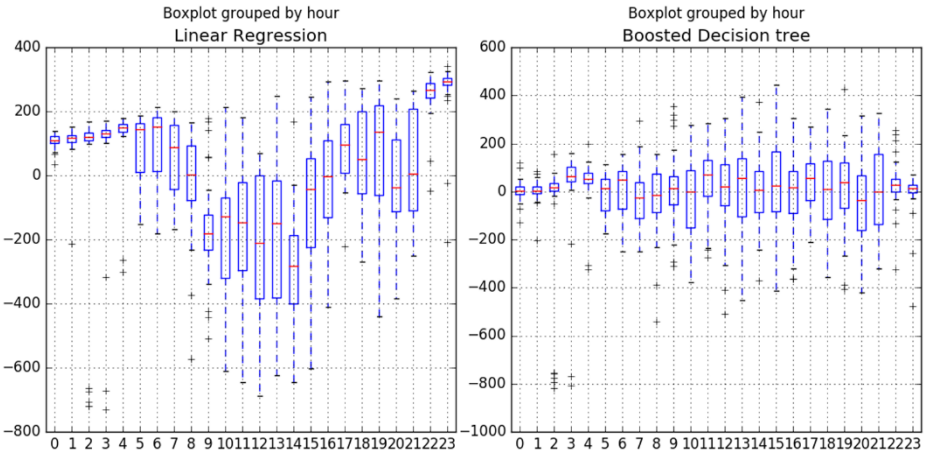

Fig. 6. Energy consumption boxplot of residuals grouped by hour.

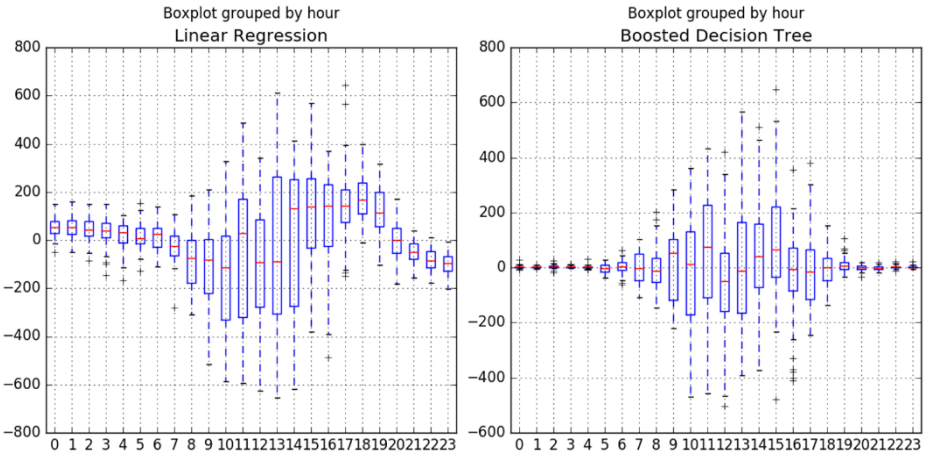

Fig. 7. PV generation boxplot of residuals grouped by hour.

The residuals are grouped by hours, across 150 days of data. As expected, it is easier to predict the end of the night and beginning of morning, when house's activity is lower and there is no solar radiation. Higher pattern changes mean higher errors.

6 Conclusions and Future Work

This work has demonstrated the application of distinct machine learning methods applied to PV energy production and energy consumption predictions, achieving 9.99% and 2.68% of mean error respectively, considering the best case (boosted decision tree). The dataset pattern is unique, so different ML methods should be applied in order to find the best one that suits each specific application.

Due to limitations of our dataset, only two houses in Ireland were analyzed. For future work the dataset will be improved adding houses, hence more historical data. Furthermore, other ML techniques could be applied, examples include Support Vector Machine and Multi-Layer Perceptron neural network. Finally, energy production from other renewables sources and storage systems can be included, adding more complexity to the proposed model.

Acknowledgements

This research work was funded by the European Union under the RESPOND project with Grant agreement No. 768619.

References

1. Communication from the Commission to the European Council and the European Parliament - an energy policy for Europe. (2007).
2. Communication from the commission to the European parliament, the council, the European economic and social committee and the committee of the regions A policy framework for climate and energy in the period from 2020 to 2030. (2014)
3. N. Shaukat, S.M. Ali, C.A. Mehmood, B. Khan, M. Jawad, U. Farid, Z. Ullah, S.M. Anwar, M. Majid, A survey on consumers empowerment, communication technologies, and renewable generation penetration within Smart Grid, Renewable and Sustainable Energy Reviews, Volume 81, Part 1, 2018, Pages 1453-1475, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2017.05.208.
4. Kadir Amasyali, Nora M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renewable and Sustainable Energy Reviews, Volume 81, Part 1, 2018, Pages 1192-1205, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2017.04.095.
5. A.S. Ahmad, M.Y. Hassan, M.P. Abdullah, H.A. Rahman, F. Hussin, H. Abdullah, R. Saidur, A review on applications of ANN and SVM for building electrical energy consumption forecasting, Renewable and Sustainable Energy Reviews, Volume 33, 2014, Pages 102-109, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2014.01.069.
6. International Energy Agency (IEA). Renewables 2017: Analysis and Forecasts to 2022, 2017. http://dx.doi.org/10.1787/re_mar-2017-en.
7. Sareh Naji, Afram Keivani, Shahaboddin Shamshirband, U. Johnson Alengaram, Mohd Zamin Jumaat, Zulkefli Mansor, Malrey Lee, Estimating building energy consumption using extreme learning machine method, Energy, Volume 97, 2016, Pages 506-516, ISSN 0360-5442, https://doi.org/10.1016/j.energy.2015.11.037.

8. Hyun Chul Jung, Jin Sung Kim, Hoon Heo, Prediction of building energy consumption using an improved real coded genetic algorithm based least squares support vector machine approach, Energy and Buildings, Volume 90, 2015, Pages 76-84, ISSN 0378-7788, https://doi.org/10.1016/j.enbuild.2014.12.029.
9. Power Technology. The True Cost of Energy Storage, 2016. http://www.power-technology.com/features/featurethe-true-cost-ofenergy. last accessed 2019/10/12
10. Mauro Castelli, Leonardo Trujillo, Leonardo Vanneschi, Aleš Popovič, Prediction of energy performance of residential buildings: A genetic programming approach, Energy and Buildings, Volume 102, 2015, Pages 67-74, ISSN 0378-7788, https://doi.org/10.1016/j.enbuild.2015.05.013.
11. Utpal Kumar Das, Kok Soon Tey, Mehdi Seyedmahmoudian, Saad Mekhilef, Moh Yamani Idna Idris, Willem Van Deventer, Bend Horan, Alex Stojcevski, Forecasting of photovoltaic power generation and model optimization: A review, Renewable and Sustainable Energy Reviews, Volume 81, Part 1, 2018, Pages 912-928, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2017.08.017.
12. Cyril Voyant, Gilles Notton, Soteris Kalogirou, Marie-Laure Nivet, Christophe Paoli, Fabrice Motte, Alexis Fouilloy, Machine learning methods for solar radiation forecasting: A review, Renewable Energy, Volume 105, 2017, Pages 569-582, ISSN 0960-1481, https://doi.org/10.1016/j.renene.2016.12.095.
13. G. Graditi, S. Ferlito, G. Adinolfi, Comparison of Photovoltaic plant power production prediction methods using a large measured dataset, Renewable Energy, Volume 90, 2016, Pages 513-519, ISSN 0960-1481, https://doi.org/10.1016/j.renene.2016.01.027.
14. Yvonne Gala, Ángela Fernández, Julia Díaz, José R. Dorronsoro, Hybrid machine learning forecasting of solar radiation values, Neurocomputing, Volume 176, 2016, Pages 48-59, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2015.02.078.
15. Aoife M. Foley, Paul G. Leahy, Antonino Marvuglia, Eamon J. McKeogh, Current methods and advances in forecasting of wind power generation, Renewable Energy, Volume 37, Issue 1, 2012, Pages 1-8, ISSN 0960-1481.
16. Ma Lei, Luan Shiyan, Jiang Chuanwen, Liu Hongling, Zhang Yan, A review on the forecasting of wind speed and generated power, Renewable and Sustainable Energy Reviews, Volume 13, Issue 4, 2009, Pages 915-920, ISSN 1364-0321.
17. RESPOND project. http://project-respond.eu/. last accessed 2019/10/12.
18. Microsoft Azure Machine Learn Studio. https://azure.microsoft.com/. last accessed 2016/10/12.
19. Pierluigi Siano, Demand response and smart grids—A survey, Renewable and Sustainable Energy Reviews, Volume 30, 2014, Pages 461-478, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2013.10.022.
20. MICROSOFT, Machine Learning studio. https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/boosted-decision-tree-regression. last accessed 2019/10/12.
21. MICROSOFT, Machine Learning studio. https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/linear-regression. last accessed 2019/10/12.
22. MICROSOFT, Machine Learning studio. https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/decision-forest-regression. last accessed 2019/10/12.
23. MICROSOFT, Machine Learning studio. https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/neural-network-regression. last accessed 2019/10/12.
24. SKYMIND website. last accessed 2019/10/12.
25. Vanderveken, B., and J. Trzcinski. "Demand Response: A study of its potential in Europe." SIA Pertners SAS, Insight Energy & Environment (2014).